Data Analysis and Prediction of Power Generated by Photovoltaic Systems

Haidar Almohri , Chongxiao Du , Zupan Hu, Jingxing Wang

Abstract— **The increasing demand for renewable energy from sources such as wind and solar has attracted the researchers to study the behavior and attributes of these energy resources in more depth. One of the important aspects of renewable energy resources is their uncertainty/unpredictability. To have a balance between the power demand and generation, it is important to know how much power could be generated in the grid at any time to avoid shortage/loss in the grid. The power generated by solar arrays mainly depends on the availability of the solar radiation (beside other factors). Since the solar radiation is unpredictable and depends on the weather condition, prediction of the output power for the solar arrays is desirable. In this work, a comprehensive analysis of the time series data as well as prediction of the output power using different regression techniques is performed. The data is collected from the NCRC solar array installed in Ann Arbor, MI. A novel approach that model by combines time series data and linear regression is developed and is found to produce the best result with the lowest error. The proposed linear regression model uses the observed values of the output as one of the predictors, along with other selected features (e.g., temperature) to predict the output.**

Index Terms— **Data Mining, Regression Modeling, Solar Energy, Time Series Analysis**

I. Introduction

In the past few decades, developed nations worldwide have widely adopted large-scale photovoltaic systems for power generation. The advantages of employing photovoltaic plants for generating electricity include no production of pollutants during operation, absence of noise pollution, long lifetime and low maintenance. Besides, solar energy is abundant, free, clean and inexhaustible [1].

These energy resources become more important when they are connected to the power grid and contribute to the power generated by the grid. The regular power generation plants such as natural gas and coal-fired power plants can manage their power generation with relative ease, usually by simply turning on/off individual units. However, the power generated by renewable energy resources such as solar and wind depends on the amount of solar insulation/wind available at the time. Since these are uncertain factors, the power output of these resources also experiences uncertainty. Furthermore, ability to estimate the amount of power that can be generated by these resources is important for the investors who plan on building and adding such resources to the grid.

The solar energy production is mainly influenced by the solar insulation, which is the amount of solar radiation energy received at a particular surface on the earth. However, other factors such as weather temperature, array temperature, and humidity may affect the power output.

In this work, we have used the data provided by the DTE Company for the solar array installed in Plymouth Rd, Ann

Arbor, MI, next to the University of Michigan north campus. DTE provided the data to the University of Michigan Energy Institute. The data includes: ambient insolation, ambient temperature, fixed solar insolation, fixed array temperature, wind speed, and the generated output power. The data is collected every 15 minutes, and is available from January 11th, 2013 up to date (we used the data until December 2013).

II. Data pre-processing

As described above, the data became available from DTE using devices installed in the location of the array that measure different factors (e.g., temperature, wind speed, etc.). Unfortunately, data was missing in between for some period, which was as long as a month and a half. Therefore, an efficient imputation method is required to complete the dataset.

A. Matrix Completion by Singular Value Thresholding

Candes and Recht (2009) proved that a low-rank, nxn matrix with m observed entries could be fully recovered with high probability, by solving a convex optimization problem if the following inequality holds true:

$$m \geq Cn^{1.2} r \log n$$

where C is a constant and r is the rank of the to-be-recovered matrix [2]. The algorithm states that for a matrix $\boldsymbol{M}$, if the set of observed entries is denoted as Ω $\{(i,j) \in \Omega$ if $\boldsymbol{M}_{ij}$ is observed$\}$, then $\boldsymbol{M}$ is recovered by solving the convex optimization problem:

$$\begin{array}{ll} \text{minimize} & \| \boldsymbol{X} \|_* \\ \text{subject to} & \boldsymbol{X}_{ij} = \boldsymbol{M}_{ij} (i,j) \in \Omega \end{array}$$

where $\boldsymbol{X}$ is the recovered matrix, and $\| \boldsymbol{X} \|_*$ is the nuclear

Manuscript received February 27, 2014; revised June 3, 2014. This work was supported by IOE department at the University of Michigan, Ann Arbor MI.

Haidar Almohri was a graduate student at the University of Michigan, Ann Arbor MI. He is now a PhD candidate at Wayne State University, Detroit MI (Phone: 413-426-1183, email:almohri@umich.edu). Chonxiao Du is a graduate student at the University of Michigan, Ann Arbor MI (email:cxdusa@umich.edu). Zupan Hu is a PhD candidate at the University of Michigan, Ann Arbor MI (email:zupanhu@umich.edu). Jingxing Wang is a graduate student at the University of Michigan, Ann Arbor MI (jeffwjx@umich.edu).

norm of the matrix ***M*** (sum of its singular values).

Since our dataset consisted of a large number of entries, and they are not correlated, it is a valid assumption to say that the dataset is a low-rank matrix. We applied the above algorithm using Matlab and a convex optimization solver package called cvx (from http://cvxr.com/cvx), to complete the dataset by imputing the missing values.

III. Feature Selection

One of the most important steps in any machine learning application is proper feature selection, as the complexity of the model as well as the accuracy of the prediction depends on this step. There are five features available in our data: (1) ambient insolation, (2) ambient temperature, (3) fixed solar insolation, (4) fixed array temperature, and (5) wind speed.

A. Feature Selection using Backward Stepwise Selection

This algorithm starts with a full model and performs an F-test in each iteration to eliminate the predictor with the smallest *F* value, and stops when removing a predictor produces an F-Statistic greater than $F_{1,N\text{-}k\text{-}1(\alpha)}$, for a predefined confidence interval (α). The F-test follows the equation :

$$F = \frac{RSS_{k-1} - RSS_k}{\frac{RSS_k}{N-k-1}}$$

where k is the kth predictor and N is the sample size.

TABLE I

RESULT OF F TEST FOR BACKWARD SUBSET SELECTION

Iteration	Predictors	Minimum F	Fmin>$F_{95\%}$?
1	[1, 2, 3, 4]	0.734	No
2	[1,2, 3]	3.842	Yes

Using this algorithm, with α = 95%, we obtained the results shown in Table 1. As is indicated in Table 1, 4 out of 5 predictors should be selected: (1) ambient insolation, (2) ambient temperature, (3) fixed solar insolation, and (4) fixed array temperature.

B. Feature Selection using LASSO (Least Absolute Shrinkage and Selection Operator)

Introduced by Robert Tibshirani (1996), LASSO minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant [3]. Because this algorithm produces some coefficients that are exactly zero, it is one of the most popular methods used for subset selection. By choosing the appropriate λ (the regularizer parameter), the same set of features was proved to be sufficient for this project.

IV. Modeling

After data pre-processing and feature selection, different algorithms are implemented and the results are analyzed. In all the following algorithms, 75% of the data is used for training, 10% for validation, and 15% for testing.

A. Combining Linear Least Square Regression and AR(2) (LLAR)

Because of the fact that the output variable is influenced by different factors (i.e., insolation and temperature), a pure time series analysis technique would fail to take these factors into consideration.

To take these factors into consideration and simultaneously take advantage of the time series analysis, we combined the linear least square regression and AR(2) time series model. In this case, the output y is a function of four predictors

x_1, x_2, x_3, x_4 as well as y_{t-1}, y_{t-2}:

$$y = f(x_1, x_2, x_3, x_4, y_{t-1}, y_{t-2})$$

Using this function, a model is fit to the data using linear least square regression:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_X^T X_t + a_t$$

where β_0, β_1, β_2, β_X are constant parameters and X_t is the design matrix that holds the features in time t. Figure 1 shows the result obtained after running the above algorithm.

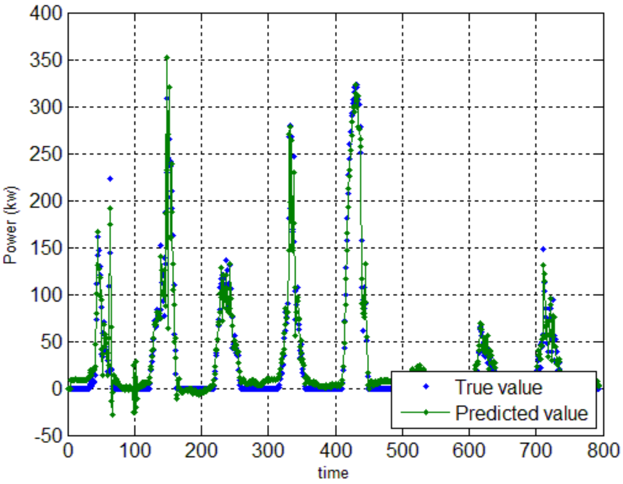

Fig 1- Result of combining linear least square regression and AR (2)

B. Kernel Ridge Regression (KRR)

A kernel ridge regression model is used to capture the nonlinearity of the data. The goal in using a kernel function is to map the data to higher dimension and use the training data to build a model and predict the output $\hat{Y}_t$ using the current features X_t as shown in the following equation:

$$\hat{Y}_t = y(K + \lambda I)^{-1}\tilde{k}(X_t)$$

where y is the row vector of the training data Y1, Y2, ..., Yn, and

$$K = \left[k(X_i, X_j)\right]_{i,j=1}^{n} \in R^{n\times n}, \qquad \tilde{k}(X_t) = \begin{pmatrix} k(X_t, X_1) \\ \vdots \\ k(X_t, X_n) \end{pmatrix} \in R^n$$

A second order polynomial kernel is used as $k(u,v)=(u^Tv+1)^2$ with λ=10. The result of this model is shown in figure 2.

C. Radial Basis Neural Network (RBNN)

Radial Basis Function (RBF) network is an artificial neural network that uses radial basis functions as activation functions:

$$r(i) = e^{(\frac{||x(i)-x||^2}{\sigma})}$$

This means that for each query point x, only the neighborhood points affect the result. In other words, the closer the point i.e. time lag, the more influence the point has on the result.

The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control [4]. This algorithm is implemented and applied to our data and the result is shown in figure 3. There are two layers in the network. The first layer is composed of radial basis neurons. The second layer is composed of linear neurons with biases. The key is to find adequate number of neurons and proper parameters for the radial basis neurons to guarantee that there is neither overfitting nor underfitting prediction. Using cross validation, 40 neurons in total (10 radial basis ones and 40 linear ones), with σ = 200 is found to produce the best result.

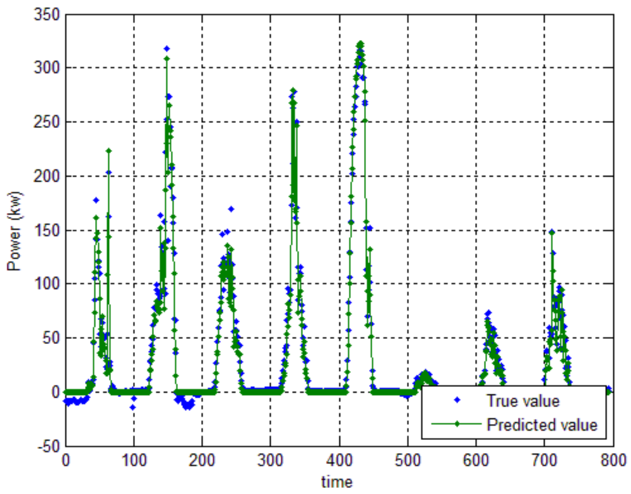

Fig 2- Result of applying kernel ridge regression

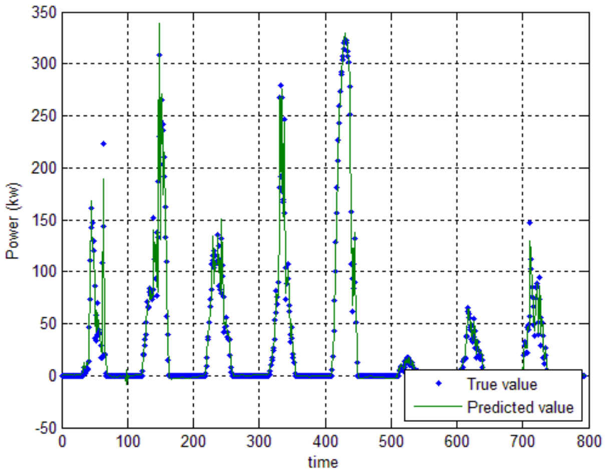

Fig 3- Result of applying Radial Basis NN

V. Results and Discussion

Table 2 summarizes the results obtained using the models introduced in previous section.

These results are obtained using 75% of the data for training, 10% for validation, and 15% for test purposes. As we can see in table 2, the "Combining Linear Least Square Regression and AR(2)" produced the least Mean Squared Error (MSE), followed by Kernel Ridge Regression, and Radial Base Neural Network. Analyzing the result, it is noticed that the MSE for the test data is larger than that of training (considering the size of each set). This can be because the models fail to produce a good result when introduced to new input. This deficiency can be improved by training the model with more data (preferably for a whole year) to reduce the variation of the result.

Table II

Summary of the Result

	MSE Training	MSE Validation	MSE Testing
LLAR	390.8853	44.1479	249.7804
KRR	356.7699	63.4635	338.9973
RBNN	493.4915	65.7957	476.5542

Another source of uncertainty is the corrupted data in our dataset. As mentioned in section 2, the data at some period was missing and the missing values were estimated. Although this data imputation is proved to be reliable, having a complete dataset can certainly improve the models and produce better results.

VI. Conclusion and Future work

In this project, a comprehensive data analysis and forecasting of the output power generated by solar array installed in Plymouth Rd, Ann Arbor, MI is performed. Because the data was corrupted at some periods, a Singular Value Thresholding algorithm is used to impute the missing data. Next, the subset selection is performed using backward and forward selection as well as Least Absolute Shrinkage and Selection Operator (LASSO) to find the predictors that best contribute in predicting the output. Finally, three different models are fit to the data and the result is analyzed. The implemented algorithms take the four inputs (ambient insulation, fixed array insulation, ambient temperature, and fixed array temperature), and predict the generated output. This model can be used for few hours to few days forecasting of the solar output since the required input data is usually available for these time periods.

Acknowledgment

We would like to thank Professor Eunshin Byon from the University of Michigan for her support during this project.

H.A thanks Professor Ratna Babu Chinnam from Wayne State University for editing the paper and providing valuable feedback.

References

[1] Chiou-Jye Huanga. , Mao-Ting, , & Chung-Cheng Chenb, (2012). A novel power output model for photovoltaic systems. *International*

Journal of Smart Grid and Clean Energy,

[2] Candès, E., Recht, B. (2009). Exact Matrix Completion via Convex Optimization.

[3] Robert Tibshirani. (1996). Regression shrinkage and selection using lasso. *Journal of Royal of Statistical Society*, *58*(1), 267-288.

[4] *Radial basis function network*. (2013, 11 02). Retrieved from http://en.wikipedia.org/wiki/Radial_basis_function_network